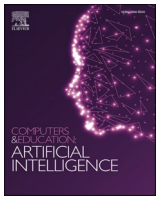
Computers and Education: Artificial Intelligence 3 (2022) 100075

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence



Assessment in the age of artificial intelligence

Zachari Swiecki [a,*], Hassan Khosravi [b], Guanliang Chen [a], Roberto Martinez-Maldonado [a], Jason M. Lodge [b], Sandra Milligan [c], Neil Selwyn [a], Dragan Gašević [a]

[a] Monash University, Australia
[b] The University of Queensland, Australia
[c] The University of Melbourne, Australia

A B S T R A C T

In this paper, we argue that a particular set of issues mars traditional assessment practices. They may be difficult for educators to design and implement; only provide discrete snapshots of performance rather than nuanced views of learning; be unadapted to the particular knowledge, skills, and backgrounds of participants; be tailored to the culture of schooling rather than the cultures schooling is designed to prepare students to enter; and assess skills that humans routinely use computers to perform. We review extant artificial intelligence approaches that–at least partially–address these issues and critically discuss whether these approaches present additional challenges for assessment practice.

1. Introduction

Well-designed assessments are essential for determining whether students have learned (Almond, Steinber, & Mislevy, 2002; Mislevy, Steinberg, & Almond, 2003). Traditional assessment practices, such as multiple-choice questions, essays, and short answer questions, have been widely used to infer student knowledge and learning (see, for example, Kaipa, 2021). In this paper, we argue that these traditional practices have several issues. First, they can be onerous for educators to design and implement. Second, they may only provide discrete snapshots of performance rather than nuanced views of learning. Third, they may be uniform and thus unadapted to the particular knowledge skills and backgrounds of participants. Fourth, they may be inauthentic, adhering to the culture of schooling rather than the cultures schooling is designed to prepare students to enter. And finally, they may be antiquated, assessing skills that humans routinely use machines to perform.

After outlining these arguments, we describe several applications of artificial intelligence (AI) that have come to–at least partially–address these issues. However, we also acknowledge that traditional assessment practices were developed for a reason and, to some extent, have been successful and valuable for understanding and improving student learning. As such, we conclude with a discussion of the unique challenges that AI may introduce to assessment practice to point to opportunities for continued research and development.

2. Background

2.1. The standard assessment paradigm

Mislevy and colleagues argue that educational assessment is often framed within the *standard assessment paradigm* (SAP) (Mislevy, Behrens, Dicerbo, & Levy, 2012). A predefined set of items (e.g., problems or questions) is used to infer claims about students' proficiency in one or more traits. The data used for these inferences are typically sparse, and student learning may not be the focus of the assessment. Instances of the SAP include widely used assessment techniques such as multiple-choice questions, essays, and short answer questions (Kaipa, 2021). While methods like these are widely used, they have several potential problems.

The first problem is a practical one. Assessments in the standard paradigm can be *onerous*. Assessment design requires carefully crafted items and techniques for translating student responses into evaluations of performance or learning—things like rubrics, answer keys, and, increasingly, sophisticated statistical models (Mislevy et al., 2012). Assessment is only one part of an educator's practice in classroom contexts. They also plan and lead learning activities, provide feedback, and, more generally, manage the classroom culture. Depending on the number of students, the other responsibilities of the educator, and how much help they have, manually designing assessments and making

* Corresponding author.
E-mail address: zach.swiecki@monash.edu (Z. Swiecki).

https://doi.org/10.1016/j.caeai.2022.100075
Received 9 August 2021; Received in revised form 4 March 2022; Accepted 21 April 2022
Available online 9 May 2022

inferences from them can be burdensome and potentially error-prone (Suto, Nádas, & Bell, 2011).[1]

Second, these assessments may be *discrete*, providing only snapshots of what students can do at a single point in time. While these snapshots may tell us something about what students do and do not know at a given time, they may tell us nothing about learning. As others have argued, one goal of assessment practices is to foster learning (see, for example, Wiliam, 2011). As understood in the learning sciences, learning is defined by *change*. For example, a change in mental representations (Perret-Clermont, 1980), a change from what you can do with help to what you can do alone (Vygotsky & Cole, 1978), a process of acclimating to a new culture (Lave & Wenger, 1991). Without comparing snapshots across time, we have no sense of change and thus no sense of learning. This logic underlies many basic analyses of learning that control for prior knowledge. Just as we would be dubious of results that report only post-tests and claim that learning was observed, we should be wary of assessments that do the same.

Relatedly, there has been a shift in the literature on learning, particularly in the learning sciences and computer-supported collaborative learning, that argues that learning processes, in addition to learning outcomes, are worthy objects of study (Puntambekar et al., 2011). Increasingly, it is becoming evident that understanding learning processes over time is critical to both student progress and fundamental questions of how learning happens (Lodge, 2018). The capacity for students to engage in effective self-regulation of their learning (e.g., Panadero, 2017), to make sound judgements about their progress (e.g., Boud, Ajjawi, Dawson, & Tai, 2018), and to change strategies when needed (e.g., Alter, Oppenheimer, Epley, & Eyre, 2007) are vital, not only for the task at hand but for longer-term learning and development of the learner. Moreover, understanding processes that are indicative or predictive of learning can help to inform feedback, interventions, and other pedagogical moves that might positively affect learning (Puntambekar et al., 2011).

Third, assessments in the SAP may be *uniform* in the sense that the same tasks or items are given to each student regardless of their prior knowledge, abilities, experiences, and cultural backgrounds. This issue is related to the first. If the assessment practice is not calibrated to the students' current state, then it speaks only to performance at the moment and not learning as we have come to define it. Moreover, viewing assessments as one-size-fits-all may introduce bias to the assessment in the sense that all students may not have equal opportunities to demonstrate their learning (Gipps & Stobart, 2009).

Fourth, assessments in the SAP are often *inauthentic*. Take essay-based assessments as an example. People for whom writing is a part of their profession write with help. They research and use the ideas of others, share drafts, get feedback, and revise; they use tools like word processors that correct their spelling, grammar, and usage, and sometimes suggest text. In contrast, writing for assessments may look quite different. Graduate study admissions tests such as the Graduate Record Examinations (GRE) ask people to write in isolation and without access to tools that have are now a standard part of writing practice (ETS, 2022). This misalignment between authentic practice and classroom culture bears on assessment more broadly. As Brown, Collins, and Duguid argue:

> When authentic activities are transferred to the classroom, their context is inevitably transmuted: they become classroom tasks and part of the school culture. Classroom procedures, as a result, are then applied to what have become classroom tasks. The system of learning and using (and, of course, *testing*) thereafter remains hermetically sealed within the self-confirming culture of the school. Consequently, contrary to the aim of schooling, success within this culture often has little bearing on performance elsewhere (Brown, Collins, & Deguid, 1989, pg. 36).

Finally, assessments in the SAP are often *antiquated* because they assess skills that are becoming increasingly obsolete. As Shaffer and Kaput (1998) argue, computational media like computers make it possible to externalise information processing much like written records make it possible to externalise information storage. This change distributes some cognitive tasks onto the computational media, for example, calculations in the case of doing mathematics with a calculator and editing in the case of writing with a word processor and frees humans up to do other kinds of tasks. These other tasks might include understanding the problem, representing the problem in a variety of external processing systems, and using the results of these systems in meaningful ways rather than doing the actual processes themselves. Consequently, they argue that, in many cases, pedagogy and–we contend—assessment should focus on the new kinds of tasks and skills afforded by external processing systems.

Despite SAPs often being discrete, uniform, isolated, and antiquated, they remain persistent in the culture of education. However, new advances in technology and artificial intelligence (AI) have come to permeate many aspects of human life—from how we work, to the products we buy, to how we spend our free time. Some classrooms as well have come to use AI as part of their everyday practice (Hwang, Xie, Wah, & Gašević, 2020). This includes relatively established technologies such as automated essay grading software (Ke & Ng, 2019) and adaptive testing (van der Linden & Glas, 2010), alongside the more recent development of continuous data-driven assessment of students' online engagements with learning materials (Shute & Rahimi, 2021). There is also increasing interest in how AI-driven monitoring and manipulation of students' engagements with online learning environments such as games and simulations can support authentic assessment of skills and behaviours exhibited in situ. In short, as Cope and colleagues argue:

> Assessment is perhaps the most significant area of opportunity offered by artificial intelligence for transformative change in education. However, this is not assessment in its conventionally understood forms. AI-enabled assessment uses dramatically different artifacts and processes from traditional assessments … Indeed, AI could spell the abandonment and replacement of traditional assessments, and with this a transformation in the processes of education (Cope, Kalantzis, & Searsmith, 2021, pg. 5).

In the following sections, we review some existing AI approaches that may help to address the issues associated with the assessment in the SAP.[2]

3. Artificial intelligence for assessment

3.1. From onerous to feasible

AI-based techniques have been developed to fully or partially automate parts of the traditional assessment practice. AI can generate assessment tasks, find appropriate peers to grade work, and automatically score student work. These techniques offload tasks from humans to AI and help to make assessment practices more feasible to maintain.

[1] Of course, some SAP instances have widely implemented automated methods to make the assessment practice less onerous. These include relatively basic methods such as the automatic scoring of multiple-choice questions and more sophisticated techniques for generating and selecting items, scoring ill-formed and open-ended responses, and making inferences from log data. We describe these techniques in relation to artificial intelligence and assessment below.

[2] Our review here is not meant to be exhaustive, but instead to highlight some exemplar approaches that we argue can address some of the existing issues with traditional assessment practice.

3.1.1. Automated assessment construction

One of the critical components of assessment design is the task used to elicit evidence to support claims about learning. In recent years, a handful of studies have been proposed to apply AI techniques to automate the generation of such assessment tasks, such as multiple-choice questions and open-answer questions. Typically, these studies are built upon AI techniques driven by deep neural networks. For instance, Jia, Zhou, Sun, and Wu (2020) proposed to improve the quality of the generated questions in a two-step manner: the representation of input text is derived by applying a Rough Answer and Key Sentence Tagging scheme, and then the input representation is further used by an Answer-guided Graph Convolutional Network to capture the inter-sentences and intra-sentence relations for question generation.

The success of such approaches often relies on the availability of large-scale and relevant datasets used to train those deep neural network models. When using these datasets to train a question generator, the source document related to each question (e.g., the transcript of a lecture video or a piece of reading material) often contains multiple sentences, and not every sentence is question-worthy. This suggests that the question-worthy sentences in an article should be first identified before we use them as input to the question generator. Driven by these findings, Chen, Yang, and Gasevic (2019) investigated the effectiveness of a total of nine sentence selection strategies in question generation and found that the stochastic graph-based method, LexRank, gave the most robust performance across multiple datasets.

While automated question generation can be a powerful tool for making assessment design more feasible for educators, it is not without its limitations. Large-scale datasets are needed to train the models that generate the questions. However, to our knowledge, most of the existing datasets are not of direct relevance to teaching and learning, except for RACE (Lai, Xie, Liu, Yang, & Hovy, 2017) and LearningQ (Chen, Yang, Hauff, & Houben, 2018). While metrics do exist for evaluating the quality of the tasks in terms of overlap between the generated questions and the human-crafted questions, for example (see Bleu-N (Papineni, Roukos, Ward, & Zhu, 2002) and Meteor (Denkowski & Lavie, 2014)) these metrics do not guarantee the pedagogical value and appropriateness of the generated questions (Horbach, Aldabe, Bexte, de Lacalle, & Maritxalar, 2020).

3.1.2. AI-assisted peer assessment

The role of high-quality feedback in learner outcomes is well attested in educational research (, in pressCarless). However, as class sizes increase, it becomes more challenging for instructors to provide rich and timely feedback. Peer assessment has been recognised as a sustainable and developmental assessment method that can address this challenge. Not only does it scale well to large class sizes, such as those in massive open online classes (MOOCs) (Shnayder & Parkes, 2016), it has also been demonstrated to promote a higher level of learning compared to one-way instructor assessment (Er, Dimitriadis, & Gašević, 2020). A range of educational platforms such as Mechanical TA (Wright, Thornton, & Leyton-Brown, 2015), Dear Beta and Dear Gamma (Glassman, Lin, Cai, & Miller, 2016), Aropä (Purchase & Hamer, 2018), CrowdGrader (De Alfaro & Shavlovsky, 2014), and RiPPLE (Khosravi, Kitto, & Williams, 2019) have been developed to support peer assessment.

Although some prior work has reported on learners' ability to evaluate resources effectively (Abdi, Khosravi, Sadiq, & Demartini, 2021; Whitehill, Aguerrebere, & Hylak, 2019), the judgements of students as experts-in-training cannot wholly be trusted, which compromises the reliability of peer assessment as an assessment instrument. However, some steps can be taken to increase reliability. One common strategy, which is used in most of the platforms mentioned above, is to rely on the wisdom of a crowd rather than one individual by employing a redundancy-based strategy and assigning the same task to multiple users. This raises a new problem commonly referred to as the consensus problem: in the absence of ground truth, how can we optimally integrate the decisions made by multiple individuals towards an accurate final decision (Zheng, Li, Li, Shan, & Cheng, 2017)?

A simple approach would be to use summary statistics such as mean or median. However, summary statistics suffer from the assumption that all students have a similar judgmental ability, which has proven incorrect (Abdi et al., 2021). An alternative is to use advanced consensus approaches that incorporate AI models to infer the reliability of each assessor (Darvishi, Khosravi, & Sadiq, 2020, 2021). Using such models allows the system to use a weighted aggregation that emphasises the marks provided by the more reliable students. A related line of research has focused on developing spot-checking methods (Wang, An, & Jiang, 2018) that optimally utilise the minimal availability of instructors to review the most controversial cases (i.e., those with low algorithmic confidence or low inter-rater agreement) and provide explanations of the outcome to learners so that they can receive valuable individualised feedback.

3.1.3. Writing analytics

The automated assessment of student writing has been a rich area of research since at least 1966 (Page, 1966). While both long-form and short answer responses have been investigated, the most successful approaches have focused on scoring longer student works. For example, several systems have been developed and used in practice for automated essay scoring, among which *MI Write* (Graham, Hebert, & Harris, 2015) is a representative.

MI Write offers a web-based interactive system for students to practice and improve their writing skills. For every essay, *MI Write* provides a student with an overall score for the essay and six trait scores (i.e., development of ideas, organisation, style, word choice, sentence fluency, and conventions) for the student to focus on specific aspects of the essay. Several studies have demonstrated that automated essay scoring tools like *MI Write* can help students to improve their writing motivation (Wilson & Czik, 2016), writing self-efficacy (Wilson & Roscoe, 2020) and writing skills (Palermo & Thomson, 2018), and help teachers to facilitate their practices and effectively influence students' writing motivation and independence (Wilson et al. (2021).

A useful survey of automated essay scoring was provided by Ke and Ng (2019), who describe the various types of AI techniques developed and applied to the problem. Typically, these AI techniques tackled the scoring task as (a) a regression task, which aimed to directly predict a score of an essay and often employed techniques like linear regression (Crossley, Allen, Snow, & McNamara, 2015) and support vector regression (Klebanov, Madnani, & Burstein, 2013); (b) a classification task, which aimed to classify an essay to one of a number categories (e.g., low quality vs. high quality) and often employed techniques like Bayesian network classification (Rudner & Liang, 2002); and (c) a ranking task, which aimed to compare essays according to their quality and often employed techniques like support vector machines (Yannakoudakis & Briscoe, 2012) and LambdaMART (Chen & He, 2013). Other tools focus more on providing feedback to students rather than an overall evaluation. For example, the tool *AcaWriter* combines natural language processing and pattern matching to identify the presence and absence of certain rhetorical moves and provide relevant feedback (Knight et al., 2020).

Another research line closely related to automated essay scoring is plagiarism detection software, e.g., *Turnitin* (Heckler, Rice, & Hobson Bryan, 2013). Different from systems used for automated essay scoring, *Turnitin* aims to compare a submission from a student against a large collection of relevant documents, which may consist of submissions from other students, online articles, and academic publications. By comparison, *Turnitin* generates a report to indicate whether there is any significant chunk of text from the submission that matches another source, which instructors can use to determine whether it is a plagiarism case. A recent systematic literature review (Foltýnek, Meuschke, & Gipp, 2019) showed significant advancement in plagiarism detection with the increased use of AI techniques - specifically, semantic text analysis methods (e.g., latent semantic analysis and word embeddings) and

machine learning algorithms.

3.2. From discrete to continuous

While traditional assessment practices may take discrete snapshots of performance, several AI techniques have been developed that afford a more continuous view of performance and thus insights into learning. Some of these approaches take traditional assessment practises such as quizzes and exams and move them to digital environments, while others apply to quite different assessment tasks and evidence.

3.2.1. Electronic assessment platforms

In recent years, *electronic assessment platforms* (EAPs) that provide the ability for exams to be administered on or off-line have become increasingly popular (Llamas-Nistal, Fernández-Iglesias, González-Tato, & Mikic-Fonte, 2013). Key advantages of EAPs include providing the ability to deliver questions that would be difficult or impossible to deliver on paper—such as questions incorporating multi-media—presenting questions in a predetermined or random order, as well as the ability to provide learners with rapid and personalised feedback (Dennick, Wilkinson, & Purcell, 2009).

As EAPs have evolved, the data extracted from each exam episode has become more sophisticated, allowing for scrutiny beyond traditional techniques like item analysis. These data may include timestamps for every action and response made by an examinee throughout their exam. Not only can these snapshots be used for exploring software bugs and investigating suspected academic misconduct, but they increasingly are used to better understand learners' behaviour. In particular, previous research has investigated: measuring and classifying test-taking effort (Wise & Gao, 2017); answering and revising behaviour during exams (Pagni et al., 2017); metacognitive regulation of strategy and cognitive processing (Goldhammer et al., 2014); the validation of test score interpretation (Engelhardt & Goldhammer, 2019); detecting rapid-guessing and pre-knowledge behaviours (Toton & Maynes, 2019); modelling examinees' accuracy, speed, and revisits (Bezirhan, von Davier, & Grabovsky, 2021); modelling students in real-time while taking a self-assessment (Papamitsiou & Economides, 2017); and understanding students' performance in various contexts such as complex problem solving (Greiff, Stadler, Sonnleitner, Wolff, & Martin, 2015).

3.2.2. Stealth assessment

Relatedly, *stealth assessment* techniques collect data that go beyond whether students have simply answered questions correctly. The term "stealth assessment" was coined by Shute and Ventura (2013) for an approach in which they used data automatically collected from learners as they played a digital game. They developed measures of conscientiousness, creativity, and physics ability by collecting data generated in a digital physics game commonly used in schools. They built models of the expected trajectory of behaviour evident in the game as students increased in capability, called a construct map (Wilson, 2005). The data were then used to place each learner on this map, generating a dynamic assessment of the increasing capability of the learner as they played.

As it was initially conceived, stealth assessment has four critical components: (a) evidence-centered assessment design (Mislevy et al., 2003), (b) formative assessment and feedback to support learning, (c) the support of pedagogical decisions, and (d) the use of learner models that may include cognitive or non-cognitive information (Shute, 2011). Typically, stealth assessment following Shute's paradigm involves unobtrusively capturing traces of learner behaviour in digital gameplay environments and modelling learners via approaches such as Bayesian networks (Pearl, 1988).

While stealth assessment refers to a specific assessment design approach, elements of it have been widely adopted in the use of digital learning environments more generally. Using similar techniques, Griffin and Care (2015) used log stream data generated from two-player digital games to assess student performance in collaborative problem-solving. Wilson and Scalise (2012) used a similar approach with log stream data generated from online tasks undertaken by students to generate measures of student ability to learn in networked digital environments. Each of these studies used custom-built digital tasks to generate the data. Milligan and Griffin (2016) extended this method to use process data derived on open platforms, using data from the log stream of MOOCs to generate assessments of learner agency. Stealth methods are now frequently used in commercial games and platforms for learning (Shute et al., 2021).

3.2.3. Latent knowledge estimation

A key component of both EAPs and stealth assessment is the ability to continuously track student actions and incorporate these actions into models of performance and learning. A widely used AI technique for generating these kinds of models is *latent knowledge estimation* (Corbett & Anderson, 1994). The reason this is referred to as *latent* lies in the fact that knowledge cannot be directly observed. What can be observed is whether a learner can apply a knowledge component in some context. This is used in intelligent tutoring systems to collect data about learners' actions to particular learning opportunities and whether they could *correctly* apply distinct knowledge components (Desmarais & Baker, 2012). This indicates that learners can produce a binary data point for each learning opportunity – they were either successful or unsuccessful in applying knowledge components.

Bayesian knowledge tracing (BKT) is the best-known technique for latent knowledge estimation (Corbett & Anderson, 1994). The technique uses four parameters to estimate whether a learner can apply a knowledge component, including (a) probability that the learner already masters a knowledge component, (b) probability of learning a knowledge component after a learning opportunity, (c) probability of correctly applying a knowledge component even when the learner has not mastered it (guess), and (d) probability of incorrectly applying a knowledge component although they know it (slip). While BKT has been widely popular, new knowledge techniques have been proposed recently based on advancements in deep learning (Gervet, Koedinger, Schneider, & Mitchell, 2020), including the use of recurrent neural networks (Piech et al., 2015) and transformers (Shin et al., 2021).

Knowledge tracing has also been used as a foundation for developing a technique – moment by knowledge learning (Baker, Goldstein, & Heffernan, 2011; 2013) – that can infer the exact moment when a learner mastered a particular skill. Not only has this technique been applied for learning about specific subject matter, but it has also been used to estimate how well learners self-regulate their learning (Molenaar, Horvers, & Baker, 2021) and offer personalised visualisations (Molenaar, 2022).

3.2.4. Learning processes

Traditional assessment practice has tended to focus on judging an artefact produced by the learner, such as an essay, a laboratory report or a completed examination sheet. The main reason it has been difficult, if not impossible, to track learning processes is that it is very time and resource-intensive. Constant monitoring of progress and the ongoing collection of indicators that allow inferences of cognitive and metacognitive processes are required. These can include self-report, behavioural, psychophysiological and other data. Collecting and analysing these data to date has been arduous, requiring specialised equipment, laboratories and analysis. Building on approaches such as stealth assessment discussed previously, AI can be used to better understand trends in learning processes.

Recent developments in multimodal data collection, learning analytics, and AI afford opportunities to improve the assessment of processes. For example, the use of multichannel data such as clickstreams, mouse movements, and eye-tracking (Azevedo & Gašević, 2019; Järvelä, Malmberg, Haataja, Sobocinski, & Kirschner, 2020) along with enhanced instrumentation of learning environments such as the use of highlights or bookmarks (Van Der Graaf et al., 2021; Jovanović, Gašević,

Pardo, Dawson, & Whitelock-Wainwright, 2019; Zhou & Winne, 2012) can offer empirical accounts about processes related to motivation, affect, cognition, and metacognition. Promising directions for assessing learning processes are being developed by analysing multichannel data with different AI and machine learning techniques such as deep learning, process mining, and network analysis (Ahmad Uzir, Gašević, Matcha, Jovanović, & Pardo, 2020; Fan, Saint, Singh, Jovanovic, & Gašević, 2021; Saint, Gašević, Matcha, Uzir, & Pardo, 2020).

3.3. From uniform to adaptive

Rather than giving the same assessment task to all students, AI techniques have been developed that adjust the task to the student's abilities, giving them tailored assessment experiences.

Computerised adaptive testing systems (CATs) conduct an exam using a sequence of successively administered questions to maximise the precision of the system's current estimate of the student's ability. There are five inter-connected technical components for building a CAT (Thompson, 2007): (1) a pool of items calibrated with pre-testing data; (2) a specific starting point for each examinee; (3) an item selection algorithm to select the next item; (4) a scoring algorithm to estimate the examinees' ability, and (5) a termination criterion for the test.

Item-response theory (IRT) (Embretson & Reise, 2013) is a common psychometric technique used in many CATs for calibrating the items. One of the key characteristics of IRT that makes it a good fit for CAT is that it places the ability of examinees and the difficulty level of items on the same metric, which helps the item selection algorithm decide which item needs to be administered next. Heuristically, an examinee is measured most effectively when test items are neither too difficult nor too easy. Given that IRT places exam-takers and items on the same metric, it can identify an item that matches the user's current ability. Consequently, if the examinee answers an item correctly, the next item selected should be more difficult; if the answer is incorrect, the next item should be easier.

To make adaptive testing operational, the size of the item pool must be large enough so that the selection algorithm can administer a suitable item based on the examinees' current ability. An important factor in a CAT is the start point. If the system has some knowledge about the examinee, it can optimise the starting point to their ability; otherwise, it may assume the examinee is of average ability. Once an item is administered, the CAT updates its estimate of the examinee's ability level. This is commonly done by updating the item response function using either *maximum likelihood estimation* and *Bayesian estimation* (Sorrel, Barrada, de la Torre, & Abad, 2020) or rating systems such as Elo rating (Abdi, Khosravi, Sadiq, & Gasevic, 2019; Verschoor, Berger, Moser, & Kleintjes, 2019). Finally, the exam is usually terminated once the system estimates the student's ability with a confidence level that exceeds a user-specified threshold. CATs have been demonstrated to have the ability to shorten the exam by 50% while maintaining higher reliability in comparison to regular exams (Collares & Cecilio-Fernandes, 2019Collares & Cecilio-Fernandes, 2019).

3.4. From inauthentic to authentic

Authentic assessments measure learning using tasks that simulate those undertaken by actual members of some community of practice (Reeves & Okey, 1996). AI techniques are now being used to augment simulated tasks and analyse the evidence associated with them.

In both virtual and physical learning environments, AI has come to play an essential role. For example, in virtual simulations called virtual internships, learners intern at a fictional company where they work in teams to design a product (Shaffer, 2006a, 2006b). The goal of virtual internships is to give learners scaffolded experience doing the kinds of things that actual professionals do, such as: conducting background research, holding design meetings, reporting to supervisors, and developing and testing prototypes. In offline simulations such as those used in healthcare, students and practitioners apply critical clinical knowledge in close-to-real-life situations (e.g., addressing an antibiotic reaction, simulating surgery) (Sullivan et al., 2018, Echeverria, Martinez-Maldonado, & Buckingham Shum, 2019). The physical learning spaces closely mimic those spaces that students will experience in the future.

Simulations for learning are designed to help learners do the kinds of things that professionals do. But in the real world it may be too difficult, expensive or dangerous to let them do so. More importantly, they necessarily lack the expertise to do so. This expertise, after all, is what they are trying to learn. To address this issue, virtual internships use AI to create an environment in which it is possible, safe, and effective for students to act like professionals. This is done via simulated professional tools, automated messages from co-workers and supervisors, and automated feedback on work products. Similarly, prospective nurses and physicians do not work with actual patients in physical healthcare simulations. In some cases, they work with simulated patients who use AI to behave like actual patients—for example, they exhibit specific symptoms at specific times (Echeverria et al., 2019).

In addition to augmenting the assessment tasks and environment, AI may collect, represent, and assess data from authentic assessments. Given that authentic assessments may involve multiple individuals or groups performing complex and ill-defined tasks, it can be challenging for educators to be aware of all that is going on during a simulation and provide detailed feedback, especially to large cohorts (Murphy, Fox, Freeman, & Hughes, 2017). Like stealth assessment and AI-driven assessments of learning processes, AI is one way to address the complexity of these assessment situations via integrated data collection and modelling.

For example, in virtual internships, the online platform automatically logs student chat messages. To relate this evidence to claims about learning, a supervised machine learning algorithm is used to automatically classify the chats as evidence of elements of an epistemic frame and epistemic network analysis (Shaffer, Collier, & Ruis, 2016) is used to identify relationships among these elements. A dashboard integrates these techniques into live representations of the epistemic networks that educators can use to monitor group interaction and plan interventions in real-time (Herder et al., 2018).

In offline simulations, multimodal learning analytics are being developed to capture millions of data points–including system logs, position coordinates, speech, and physiological traces–in physical spaces and in a relatively short amount of time. AI may be integral to the functioning of these sensors, as with the case of automated transcription tools. To make these data available for educators, one approach that has been adopted is to use *data storytelling* principles to create interfaces in which stories are extracted from the complex multimodal data to focus on one learning or reflection goal at a time. For example, Echeverria and colleagues (2020) focused on creating data stories related to common errors performed by nursing students based on the automated assessments of the sequence and timeliness of their logged actions.

3.5. From antiquated to modern

Computational media like computers, calculators, and software make it possible to externalise information processing in new and powerful ways. While computational media exist in various domains, here we briefly focus on some of those developed for writing tasks as an example.

Digital word processors have been in use since at least the 1970s (Bergin, 2006). In addition to simply recording and storing text, their primary function has been to offload typical writing tasks, such as editing, from humans to computers. Digital word processors commonly include automated techniques for checking spelling, grammar, and usage. As these tools have developed, they have increasingly come to rely on AI to complete more sophisticated tasks.

Today's digital word processors like Microsoft Word and Google

Docs include AI techniques that suggest word and sentence completions (Microsoft, 2022). Other commercial tools, like *Grammarly* (Grammarly, 2022), include AI that infers tone and style. AI-based tools like *Sudowrite* (Marche, 2021) now exist that generate entirely new sections of text based on a few sample lines. Because these tools may be used by learners and professionals in their everyday practices, assessment designs may incorporate them. Using tools to do increasingly complex and humanlike tasks has important implications for assessment, some of which we discuss below.

4. Challenges for AI and assessment

Thus far, we have highlighted a set of issues with the SAP and reviewed some AI-based approaches that bear on these issues. While the sections above suggest that AI can improve the SAP, we acknowledge that this paradigm has a long and, arguably, successful history. It is worth, then, reflecting on what we might lose–or other problems we might introduce–by introducing AI to this paradigm.

4.1. The sidelining of professional expertise

Many researchers seek to develop AI technologies that support and guide teachers' decision-making, freeing teachers from routine, uncontentious tasks and decisions while continuing to defer to teachers' ultimate judgment and oversight (see, for example, Herder et al., 2018). In this sense, it is reassuring to imagine that AI-enabled assessment will retain humans-in-the-loop, with teachers able to oversee and override any automated decision when they see fit.

However, one potential danger of automated decision-making is the sidelining of professional expertise—that is, machine calculations and outputs being deferred to or automatically taken as correct. A hypothetical example of this can be seen with plagiarism software at educational institutions. In the past, teachers made decisions regarding whether student submissions were too similar to one another or available sources. However, given the volume of possible sources and advances in natural language processing, AI can now handle this task in many contexts. Given the difficulty of this task and the efficacy of existing algorithms, it is possible, and perhaps easy, for educators to take their output as a correct decision rather than a tentative suggestion. It would take a confident and time-rich teacher to regularly challenge these systems' outputs. As such, there are understandable concerns that we face the prospect of teachers' decision-making capacity being 'hollowed out' as automated assessment systems "creat[e] a distance between their decisions and the evidence-gathering processes on which those decisions must rely" (Couldry, 2020, p. 1139).

To prevent such a hollowing-out, researchers have begun to design systems in which the decision-making processes are *explainable* to the teacher (Rosé, McLaughlin, Liu, & Koedinger, 2019; Khosravi et al., 2022). While this is a promising direction, more work is needed to better understand the balance between AI and teacher decision-making that is best for teaching, learning, and assessment.

4.2. The black-boxing of accountability

While many researchers might argue that it is not their intention to do so (see, for example, Baker, 2016), taking human teachers out of the assessment loop is likely to be an appealing prospect for many key stakeholders involved in school and university education. Educational institutions may welcome the capacity for the reliable, timely production of assessment data at scale—avoiding inconsistencies over mismarking or delays resulting from the marking simply not being done on time.

Similarly, many teachers may be happy to defer responsibility and dodge the awkward task of personally grading students that they have grown to know—particularly given current tendencies for students to appeal and contest grades, and even initiate legal action over misgrading (see, for example, Griffiths, 2021). Students too might welcome the option of not having to subject themselves to face the vulnerability of being judged by their teachers, schools or other social institutions close to home—in other words, the frictions of being assessed by people who actually know them.

Yet, AI-enabled assessment is not a simple case of deferring educational judgements to the dispassionate, objective, reliable gaze of the machine. There is no such thing as neutral, dispassionate non-human assessment (Mayfield et al., 2019; Scheuneman, 1979). Instead, AI-enabled assessment can more accurately be described as handing those decisions over to programmers, learning engineers, instructional designers, software vendors and other humans that have no direct knowledge of the students being assessed, their local contexts, or even necessarily the educational systems that they are studying within. Thus, as with any form of assessment, AI-enabled assessment is an objective-partial process. As Hanesworth and colleagues put it:

> No matter the structures and processes put in place, assessments are designed and evaluated by humans, with all their complex sociocultural backgrounds, educational experiences, and intellectual and personal values (Hanesworth, Bracken, & Elkington, 2019, pg. 99).

In the case of AI-based assessment, the responsibility for the modelling and execution of educational assessment is deferred to distant others (programmers, learning engineers). On the one hand, this can be welcomed as distancing assessment decisions from the biases and assumptions of classroom teachers. Yet, on the other hand, this also raises concerns that need to be taken more seriously in terms of how AI-enabled assessment then exposes the student to the biases, values, assumptions of those other people who otherwise have no knowledge of or personal investment in those who are being assessed.

At the very least, in practical terms, these concerns raise the pressing need for rigorous oversight of any AI-enabled assessment and the establishment of clear lines of accountability for the decisions that these systems and software produce—as well as clear lines of accountability for how software outputs are then translated over into final grades by educational institutions.

4.3. Restricting the pedagogical role of assessment

Amidst the current enthusiasm for AI-enabled assessment, there is little acknowledgement of the pedagogical role of assessment. This relates to the idea that educational assessment is not solely a matter of gauging what a student has (and has not) learnt (Wiliam, 2011). Instead, when considering the consequences of increased use of AI-based assessments, it is important to consider how this might impact the ability of educators to engage with assessment as a pedagogical act.

For example, on a personal level, teachers will often use traditional forms of teacher-graded assessment to motivate, support and cajole students (Cauley & McMillan, 2010; Harlen, 2012). This might involve showing leniency when the teacher feels that a student will benefit from being encouraged and seen to succeed. Alternatively, this might involve being more punitive where a teacher feels that a student might benefit from an intervention. In both instances, the act of assessment is rooted in the personal relationships and knowledge that a teacher has established with her student.

Many educators also pay close attention to what is learnt from any assessment act. This is implicit in some educators' use of alternate forms of assessment. For example, the rising popularity of peer assessment is rooted primarily as a means of encouraging self-reflection among students on their own work (Cho & Cho, 2011; Topping, 2018). The trend for allowing student-led self-assessment is similarly based on intentions to develop student deliberation on one's own learning practices. Similarly, growing interest in the use of 'assessment for social justice' seeks to support students' engagement with multiple and contested perspectives and dealing with variation arising from contextual differences, historical aspects and personal normativities (see McArthur, 2016,

Hanesworth et al., 2019). This might entail, for example, allowing students to take a leading role in collectively deciding on the nature and form of how they are assessed. In all cases, the intention is to support students to reflect on educational processes and practices rather than produce an objective 'measure' of learning.

Concerns can be raised that some AI-enabled assessments prevent teachers from using assessment in these alternate ways. Yet, such examples also highlight the value-driven nature of how educational assessment is undertaken—an aspect that has not featured in many discussions of AI-enabled assessment. The idea of 'assessment for social justice' certainly conveys a distinct set of values about what education is and what education is for. This, in turn raises questions about the implicit values and ideological underpinnings of AI-enable assessment. Is it fair to argue, as Saltman (2020, p.199) implies, that AI approaches to education appear to promote ideals of "standardized and transmission-oriented approaches to teaching"? Or, that AI-enabled assessment corresponds closely with the employment conditions of the post-Fordist neoliberal workplace—preparing future workers for conditions of continual tracking, monitoring of performance, nudging of behaviours, and so on. These are concerns that the community that works on AI-enabled assessment need to engage with. If not these ideals and values, then what are the values and ideological underpinnings that are being advanced through the development of AI-enabled assessment?

Researchers have also begun to address this issue, at least implicitly. For example, several researchers have called for a more prominent role for educational and learning theory in the development of AI approaches (Rogers, Gasevic, & Dawson, 2016, pp. 232–250). These theories take a stance on what is valued with respect to learning, and they may differ markedly from transmission-oriented approaches, instead, focusing, for example, on promoting the ideals of particular communities of practice (Shaffer, 2006a, 2006b) or the ability to regulate one's learning (Azevedo and Gašević, 2019; Molenaar, 2022).

4.4. Assessing limited forms of learning

Extending the idea of AI-enabled assessment as curtailing different forms of teaching are concerns over restricted forms of learning implicit in the use of AI-enabled assessment. Of course, one of the central promises of AI-enabled assessment is the capacity to recognise and respond to all the forms of learning prevalent in the digital age—to know things about what has been learnt that would otherwise remain unknown. Yet this promise of comprehensive assessment of learning in all its forms obscures that any form of assessment demarcates and delineates what is understood by learning in any education system (Messick, 1994). As Taras (2008, p.389) puts it, assessment is "the single most important component" that shapes student learning.

In this sense, concerns can be raised that many forms of AI-enabled assessment perpetuate the orientation of current traditional assessment regimes toward emphasising skills, rational thinking and behaviours, alongside predominantly white, male, middle-class, Western values of objectivity and individualism (Hanesworth et al., 2019). In other instances, the prominence of technologies such as eye-tracking highlights the dangers of AI-enable-assessment acting to reinforce and privilege ableist—and especially—neurotypical models of learning and what it means to exhibit learning-related behaviours (Swauger, 2020). All told, strong arguments can be made that AI-enabled assessment may well alter—but not necessarily expand—the forms of learning that are being assessed.

Thus, conversations in the research community need to explore the contention that AI-enabled assessment is not a neutral site where any form of learning will be detected and assessed. For example, as with any form of assessment, it could be argued that any instance of AI-enabled assessment will inevitably codify specific cultural, disciplinary and individual norms, value systems and knowledge hierarchies. Moreover, it may inculcate these norms, values and knowledge hierarchies within students. Students will learn to perform in ways that are algorithmically assessable and algorithmically rewarded. Put another way, "teaching to the test" (Popham, 2001) is not necessarily avoided using AI-enabled assessment.

At the same time, there is also a need for discussions of AI-enabled assessment to better acknowledge the many forms of learning that cannot yet be detected, measured, and modelled by non-humans. AI software is notoriously limited in detecting meaning in language or images—be it the simple development of a logical argument to nuance and inflection such as irony and sarcasm. For example, natural language processing technology might have a near-infinite capacity to recognise vocabulary but remains tone-deaf to the subtleties of language—double-meanings, allusions, local vernacular, tone and subtext.

Similarly, AI-enabled assessment may remain understandably limited in its capacity to recognise (let alone assess) instances of improvisation, creativity, poetry, morals, or ethics. There may be little room for recognising (and rewarding) distinctly different, unexpected and perhaps unique ways of setting about a learning task— where students engage in genuine originality and 'out of the box' thinking that a good human assessor would be able to appreciate (even if they would have never thought of it themselves). In short, there exist aspects of learning that remain perceptible to humans but not machines. As such, discussions of AI-enabled assessment need to be more forthcoming in acknowledging what the technology cannot (and may never) be capable of assessing.

4.5. Surveillance pedagogy

In one sense, AI-enabled assessment builds on some distinct logics of 'datafication' in education, such as the idea of continuous, comprehensive data generation relating to an individual's ongoing engagement with an online learning environment. This evokes promises of continuous assessment that are not necessarily recognised by students as assessment – thereby overcoming issues of 'test anxiety' (Colwell, 2013) and allowing for all aspects of an individual's learning to be made visible. However, these promises of continual background data monitoring can be seen to constitute conditions of surveillance. As such, the promise of data-driven educational environments "to make visible what might otherwise be hidden or missed" (Bayne et al., 2020, p. 185) needs to be acknowledged as potentially problematic, as well as potentially beneficial.

For example, there needs to be more acknowledgement in discussions of AI-enabled assessment regarding how this state of continuous surveillance also lends itself to processes of control and compliance—for example, monitoring for indications of malpractice and other forms of cheating. Of course, most aspects of formal education institutions such as schools and universities are seen to be based traditionally around 'surveillance pedagogies'—not least the traditional set-up of the classroom or the examination hall—with seats arranged in rows, facing the front of the class, teacher supervising student bodies (Luke, 2003; McLaren & Leonardo, 1998). Nevertheless, online education (and, by implication, AI-enabled assessment) extends and amplifies the scope of this surveillance to all times and all spaces of the school day or the university experience.

In this sense, it could be argued that AI-enabled assessment constitutes an administrative—rather than a pedagogic—gaze, impinging on the fragile conditions of trust that most educators see as underpinning the teacher/student relationship:

> … in higher education settings, a culture of surveillance, facilitated and intensified by technology, risks creating conditions that are highly risk averse and destructive of the trust basis on which academic and student autonomy and agency rely. Technology architectures introduced to build trust by mapping performance may end up directly undermining these very goals" (Bayne et al., 2020, p.182).

It is also important to better consider the implications of continuous

surveillance of students in terms of pedagogical lines. In particular, the SAP also conveys that learning can often best take place where there is no assessment. This contrasts with the benefits of continuous and comprehensive monitoring and assessment of students' educational engagement. As such, while by no means perfect, current forms of education are set up in ways that support learning and progression to occur during episodes where there is no assessment. Well-designed teaching offers many moments of rehearsal—recognising the vulnerability of learning and allowing students ample opportunity to learn in private, engage in preparatory work, experiment, make mistakes, and fail. In other words, the absence of assessment is seen as the best condition for learning and progression.

However, the importance of the absence of assessment also seems contradictory. How do we know whether it is good for learning if we cannot tell whether learning is occurring? Perhaps one way around this issue is to continue to shift the focus of assessment from evaluation or judgement to development. In this view, continual monitoring of student processes is not a means of determining whether someone is doing something "right" or "wrong", but instead, monitor for opportunities to provide feedback and improve learning (Wiliam, 2011). Thus, what is needed is not necessarily a shift in how the assessments take place but a conceptual shift in what they mean and what they are for.

4.6. Distributed assessment models

A final concern has to do with the changes that computational tools imply for assessment. One way to characterise assessment is as an argument from evidence (Messick, 1994). In *evidence-centered* assessment design, for example, this argument includes a *student model* that describes the traits, skills, or abilities to be assessed; a *task model* that describes activities students will do to produce evidence that they have those traits; and an *evidence model* that describes the variables and techniques that will be used to relate the evidence to the traits. One consequence of AI-based computational tools is that they complicate each of these models.

In terms of the student model, the presence of AI suggests that we should adjust the traits, skills, and abilities assessed to be those that require human influence rather than those that AI can accomplish on their own. In terms of the task model, AI suggests that we should allow students to use AI-based computational tools during the assessment. And in terms of the evidence model, the presence of AI suggests that we should account for the fact that a human-AI team can generate assessment evidence. Depending on the sophistication of the AI, this could mean trying to separate the human and AI contributions, accounting for the relationship between these contributions, or treating them as if they came from the same source. While some attempts have been made to integrate assessment design theory with AI (see Mislevy et al., 2012), to date, they have mainly focused on the applications of AI to the evidence model and less so on the task and student models.

5. Conclusion

We have argued that several issues mar the standard assessment paradigm.

First, assessments in this paradigm can be onerous for educators to design and implement. Second, they may only provide discrete snapshots of performance rather than nuanced views of learning. Third, they may be uniform and thus unadapted to the particular knowledge skills and backgrounds of participants. Fourth, they may be inauthentic, adhering to the culture of schooling rather than the cultures schooling is designed to prepare students to become members of. And finally, they may be antiquated, assessing skills that humans routinely use machines to perform.

While extant artificial intelligence approaches partially address the issues above, they are not a panacea. As our discussion highlights, these approaches bring with them a new set of challenges that must be considered when designing and implementing assessments. We hope that this paper brings both the issues with the standard assessment paradigm and the challenges associated with AI and assessment into a deeper conversation that will ultimately improve assessment practices more generally.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by funding from the Australian Research Council (DP210100060, DP220101209, CIRES/IC200100022), Economic and Social Research Council of the United Kingdom (ES/S015701/1), and Jacobs Foundation (CELLA 2 CERES, Research Fellowship Program). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors. They do not necessarily reflect the views of the funding agencies, cooperating institutions, or other individuals.

References

Abdi, S., Khosravi, H., Sadiq, S., & Demartini, G. (2021). Evaluating the quality of learning resources: A learner sourcing approach. *IEEE Transactions on Learning Technologies, 14*(1), 81–92.

Abdi, S., Khosravi, H., Sadiq, S., & Gasevic, D. (2019). A multivariate ELO-based learner model for adaptive educational systems. In *Proceedings of the 12th international conference on educational data mining* (pp. 462–467).

Ahmad Uzir, N., Gašević, D., Matcha, W., Jovanović, J., & Pardo, A. (2020). Analytics of time management strategies in a flipped classroom. *Journal of Computer Assisted Learning, 36*(1), 70–88.

Almond, R. G., Steinber, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *The Journal of Technology, Learning, and Assessment, 5*.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General, 136*, 569–576. https://doi.org/10.1037/0096-3445.136.4.569

Azevedo, R., & Gašević, D. (2019). Analyzing multimodal multichannel data about self-regulated learning with advanced learning technologies: Issues and challenges. *Computers in Human Behavior, 96*, 207–210.

Baker, R. S. (2016). Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education, 26*(2), 600–614.

Baker, R. S., Goldstein, A. B., & Heffernan, N. T. (2011). Detecting learning moment-by-moment. *International Journal of Artificial Intelligence in Education, 21*(1–2), 5–25.

Baker, R. S., Hershkovitz, A., Rossi, L. M., Goldstein, A. B., & Gowda, S. M. (2013). Predicting robust learning with the visual form of the moment-by-moment learning curve. *The Journal of the Learning Sciences, 22*(4), 639–666.

Bayne, S., Evans, P., Ewins, R., Knox, J., Lamb, J., Macleod, H., et al. (2020). *The manifesto for teaching online*. MIT Press.

Bergin, T. J. (2006). The origins of word processing software for personal computers: 1976-1985. *IEEE Annals of the History of Computing, 28*(4), 32–47.

Bezirhan, U., von Davier, M., & Grabovsky, I. (2021). Modeling item revisit behavior: The hierarchical speed–accuracy–revisits model. *Educational and Psychological Measurement, 81*(2), 363–387.

Boud, D., Ajjawi, R., Dawson, P., & Tai, J. (2018). *Developing evaluative judgement in higher education: Assessment for knowing and producing quality work*. Abingdon, UK: Routledge.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32–42.

Carless, D. (2022). From teacher transmission of information to student feedback literacy: Activating the learner role in feedback processes. *Active Learning in Higher Education*. https://doi.org/10.1177/1469787420945845 (in press).

Cauley, K. M., & McMillan, J. H. (2010). Formative assessment techniques to support student motivation and achievement. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 83*(1), 1–6.

Chen, H., & He, B. (2013). Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1741–1752).

Chen, G., Yang, J., & Gasevic, D. (2019). A comparative study on question-worthy sentence selection strategies for educational question generation. In *Proceedings of the 20th international conference on artificial intelligence in education* (pp. 59–70). Cham: Springer.

Chen, G., Yang, J., Hauff, C., & Houben, G. J. (2018). LearningQ: A large-scale dataset for educational question generation. In *Proceedings of the 12th international AAAI conference on web and social media* (pp. 481–490). AAAI.

Cho, Y. H., & Cho, K. (2011). Peer reviewers learn from giving comments. *Instructional Science, 39*(5), 629–643.
Collares, C. F.Cecilio-Fernandes, D., … (2019). When I say computerised adaptive testing. *Medical Education, 53*(2), 115–116.
Colwell, N. M. (2013). Test anxiety, computer-adaptive testing and the common core. *Journal of Education and Training Studies, 1*(2), 50–60.
Cope, B., Kalantzis, M., & Searsmith, D. (2021). Artificial intelligence for education: Knowledge and its assessment in AI-enabled learning ecologies. *Educational Philosophy and Theory*. https://doi.org/10.1080/00131857.2020.1728732
Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*(4), 253–278.
Couldry, N. (2020). Recovering critique in an age of datafication. *New Media & Society, 22*(7), 1135–1151.
Crossley, S., Allen, L. K., Snow, E. L., & McNamara, D. S. (2015). Pssst... textual features... there is more to automatic essay scoring than just you. In *Proceedings of the fifth international conference on learning analytics and knowledge* (pp. 203–207).
Darvishi, A., Khosravi, H., & Sadiq, S. (2020). Utilising learner sourcing to inform design loop adaptivity. In *Proceedings of the 14th European conference on technology-enhanced learning* (pp. 332–346). Springer.
Darvishi, A., Khosravi, H., & Sadiq, S. (2021). Employing peer review to evaluate the quality of student generated content at scale: A trust propagation approach. In *Proceedings of the eighth ACM conference on learning@ scale* (pp. 139–150).
De Alfaro, L., & Shavlovsky, M. (2014). Crowdgrader: A tool for crowdsourcing the evaluation of homework assignments. In *Proceedings of the 45th ACM technical symposium on computer science education* (pp. 415–420).
Denkowski, M., & Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation* (pp. 376–380).
Dennick, R., Wilkinson, S., & Purcell, N. (2009). Online eAssessment: AMEE guide no. 39. *Medical Teacher, 31*(3), 192–206.
Desmarais, M. C., & Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction, 22*(1), 9–38.
Echeverria, V., Martinez-Maldonado, R., & Buckingham Shum, S. (2019). Towards collaboration translucence: Giving meaning to multimodal group data. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–16).
Educational Testing Service. (2022, March 1). *What to expect during the GRE general test"*. https://www.ets.org/gre/revised_general/test_day/expect.
Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.
Engelhardt, L., & Goldhammer, F. (2019). Validating test score interpretations using time information. *Frontiers in Psychology, 10*, 1131.
Er, E., Dimitriadis, Y., & Gašević, D. (2020). A collaborative learning approach to dialogic peer feedback: A theoretical framework. *Assessment & Evaluation in Higher Education, 46*(4), 586–600.
Fan, Y., Saint, J., Singh, S., Jovanovic, J., & Gašević, D. (2021). April). A learning analytic approach to unveiling self-regulatory processes in learning tactics. In *LAK21: 11th international learning analytics and knowledge conference* (pp. 184–195).
Foltýnek, T., Meuschke, N., & Gipp, B. (2019). Academic plagiarism detection: A systematic literature review. *ACM Computing Surveys, 52*(6), 1–42.
Gervet, T., Koedinger, K., Schneider, J., & Mitchell, T. (2020). When is deep learning the best approach to knowledge tracing? *Journal of Educational Data Mining, 12*(3), 31–54.
Gipps, C., & Stobart, G. (2009). Fairness in assessment. In *Educational assessment in the 21st century* (pp. 105–118). Dordrecht: Springer.
Glassman, E. L., Lin, A., Cai, C. J., & Miller, R. C. (2016). Learner sourcing personalized hints. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing* (pp. 1626–1636).
Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608–626. https://doi.org/10.1037/a0034716
Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *The Elementary School Journal, 115*(4), 523–547.
Grammarly. (2022, March 1). *About grammarly*. https://www.grammarly.com/about.
Greiff, S., Stadler, M., Sonnleitner, P., Wolff, C., & Martin, R. (2015). Sometimes less is more: Comparing the validity of complex problem solving measures. *Intelligence, 50*, 100–113.
Griffin, P., & Care, E. (2015). *Assessment and teaching of 21st century skills: Methods and approaches* (Vol. 2). Dordrecht: Springer.
Griffiths, S. (2021). *Families to sue over 'wrong' marks given by teachers*. The Times. Retrieved from https://www.thetimes.co.uk/article/families-to-sue-over-wrong-marks-given-by-teachers-g2qjjc8x7.
Hanesworth, P., Bracken, S., & Elkington, S. (2019). A typology for a social justice approach to assessment. *Teaching in Higher Education, 24*(1), 98–114.
Harlen, W. (2012). The role of assessment in developing motivation for learning. In J. Gardner (Ed.), *Assessment and learning* (2nd ed., pp. 61–80). SAGE Publications.
Heckler, N. C., Rice, M., & Hobson Bryan, C. (2013). Turnitin systems: A deterrent to plagiarism in college classrooms. *Journal of Research on Technology in Education, 45*(3), 229–248.
Herder, T., Swiecki, Z., Fougt, S. S., Tamborg, A. L., Allsopp, B. B., Shaffer, D. W., et al. (2018). Supporting teachers' intervention in students' virtual collaboration using a network based model. In *Proceedings of the 8th international conference on learning analytics and knowledge* (pp. 21–25).
Horbach, A., Aldabe, I., Bexte, M., de Lacalle, O. L., & Maritxalar, M. (2020). Linguistic appropriateness and pedagogic usefulness of reading comprehension questions. In *Proceedings of the 12th language resources and evaluation conference* (pp. 1753–1762).
Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of Artificial Intelligence in Education. *Computers & Education: Artificial Intelligence, 1*, Article 100001.
Järvelä, S., Malmberg, J., Haataja, E., Sobocinski, M., & Kirschner, P. A. (2020). What multimodal data can tell us about the students' regulation of their learning process. *Learning and Instruction, 45*, Article 100727.
Jia, X., Zhou, W., Sun, X., & Wu, Y. (2020). *EQG-RACE: Examination-Type question generation*. arXiv preprint arXiv:2012.06106.
Jovanović, J., Gašević, D., Pardo, A., Dawson, S., & Whitelock-Wainwright, A. (2019). Introducing meaning to clicks: Towards traced-measures of self-efficacy and cognitive load. In *Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 511–520). New York: ACM.
Kaipa, R. M. (2021). Multiple choice questions and essay questions in curriculum. *Journal of Applied Research in Higher Education, 13*(1), 16–32. https://doi.org/10.1108/JARHE-01-2020-0011
Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In *Proceedings of the 28th international joint conference on artificial intelligence* (pp. 6300–6308).
Khosravi, H., Conati, C., Martinez-Maldonado, R., Knight, S., Kay, J., Chen, G., et al. (2022). Explainable AI in education. *Computers & Education: Artificial Intelligence*. In this issue.
Khosravi, H., Kitto, K., & Williams, J. J. (2019). *Ripple: A crowdsourced adaptive platform for recommendation of learning activities. arXiv preprint arXiv:1910.05522*.
Klebanov, B. B., Madnani, N., & Burstein, J. (2013). Using pivot-based paraphrasing and sentiment profiles to improve a subjectivity lexicon for essay data. *Transactions of the Association for Computational Linguistics, 1*, 99–110.
Knight, S., Shibani, A., Abel, S., Gibson, A., Ryan, P., Sutton, N., et al. (2020). Acawriter: A learning analytics tool for formative feedback on academic writing. *Journal of Writing Research, 12*(1), 141–186.
Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017). *Race: Large-scale reading comprehension dataset from examinations*. arXiv preprint arXiv:1704.04683.
Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press.
van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. New York, NY: Springer.
Llamas-Nistal, M., Fernández-Iglesias, M. J., González-Tato, J., & Mikic-Fonte, F. A. (2013). Blended e-assessment: Migrating classical exams to the digital world. *Computers & Education, 62*, 72–87.
Lodge, J. M. (2018). A futures perspective on information technology and assessment. In J. Voogt, G. Knezek, R. Christensen, & K.-W. Lai (Eds.), *International handbook of information technology in primary and secondary education* (2nd ed., pp. 1–13). Berlin: Springer.
Luke, C. (2003). Pedagogy, connectivity, multimodality, and interdisciplinarity. *Reading Research Quarterly, 38*(3), 397–403.
Marche, S. (2021, April 3). *The computers are getting better at writing*. https://www.newyorker.com/culture/cultural-comment/the-computers-are-getting-better-at-writing.
Mayfield, E., Madaio, M., Prabhumoye, S., Gerritsen, D., McLaughlin, B., Dixon-Román, E., et al. (2019, August). Equity beyond bias in language technologies for education. In *Proceedings of the 14th workshop on innovative use of NLP for building educational applications* (pp. 444–460).
McArthur, J. (2016). Assessment for social justice. *Assessment & Evaluation in Higher Education, 41*(7), 967–981.
McLaren, P., & Leonardo, Z. (1998). Deconstructing surveillance pedagogy. *Studies in the Literary Imagination, 31*(1), 127–147.
Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.
Microsoft. (2022, March 1). Microsoft Editor checks grammar and more in documents, mail, and the web. https://support.microsoft.com/en-us/office/microsoft-editor-checks-grammar-and-more-in-documents-mail-and-the-web-91ecbe1b-d021-4e9e-a82e-abc4cd7163d7.
Milligan, S. K., & Griffin, P. (2016). Understanding learning and learning design in MOOCs: A measurement-based interpretation. *Journal of Learning Analytics, 3*(2), 88–115.
Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of educational data mining, 4*(1), 11–48.
Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–62.
Molenaar, I. (2022). The concept of hybrid human-AI regulation: Exemplifying how to support young learners' self-regulated learning. *Computers & Education: Artificial Intelligence*. In this issue.
Molenaar, I., Horvers, A., & Baker, R. S. (2021). What can moment-by-moment learning curves tell about students' self-regulated learning? *Learning and Instruction, 72*, Article 101206.
Murphy, V., Fox, J., Freeman, S., & Hughes, N. (2017). Keeping it real": A review of the benefits, challenges and steps towards implementing authentic assessment. *All Ireland Journal of Higher Education, 9*(3), 1–13.
Page, E. B. (1966). The imminence of... grading essays by computer. *Phi Delta Kappan, 47*(5), 238–243.
Pagni, S. E., Bak, A. G., Eisen, S. E., Murphy, J. L., Finkelman, M. D., & Kugel, G. (2017). The benefit of a switch: Answer-changing on multiple-choice exams by first-year dental students. *Journal of Dental Education, 81*(1), 110–115.
Palermo, C., & Thomson, M. M. (2018). Teacher implementation of self-regulated strategy development with an automated writing evaluation system: Effects on the argumentative writing performance of middle school students. *Contemporary Educational Psychology, 54*, 255–270.

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology, 8*, 883–928. https://doi.org/10.3389/fpsyg.2017.00422

Papamitsiou, Z., & Economides, A. A. (2017). Student modeling in real-time during self-assessment using stream mining techniques. In *Proceedings of the 17th IEEE international conferenceon advanced learning technologies* (pp. 286–290). IEEE.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318).

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Kaufmann.

Perret-Clermont, A.-N. (1980). *Social interaction and cognitive development in children*. Academic Press.

Piech, C., Spencer, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., et al. (2015). *Deep knowledge tracing. arXiv preprint arXiv:1506.05908*.

Popham, W. J. (2001). Teaching to the test? *Educational Leadership, 58*(6), 16–21.

Puntambekar, S., Erkens, G., & Hmelo-Silver, C. (Eds.). (2011). *Analyzing interactions in CSCL*. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4419-7710-6.

Purchase, H., & Hamer, J. (2018). Peer-review in practice: Eight years of Aropä. *Assessment & Evaluation in Higher Education, 43*(7), 1146–1165.

Reeves, T. C., & Okey, J. R. (1996). Alternative assessment for constructivist learning environments. In B. G. Wilson (Ed.), *Constructivist learning environments: Case studies in instructional design* (pp. 191–202). Englewood Cliffs, NJ: Educational Technology Publications.

Rogers, T., Gašević, D., & Dawson, S. (2016). *Learning analytics and the imperative for theory driven research*. The SAGE Handbook of E-Learning Research.

Rosé, C. P., McLaughlin, E. A., Liu, R., & Koedinger, K. R. (2019). Explanatory learner models: Why machine learning (alone) is not the answer. *British Journal of Educational Technology, 50*(6), 2943–2958.

Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning, and Assessment, 1*(2).

Saint, J., Gašević, D., Matcha, W., Uzir, N. A. A., & Pardo, A. (2020). Combining analytic methods to unlock sequential and temporal patterns of self-regulated learning. In *Proceedings of the tenth international conference on learning analytics & knowledge* (pp. 402–411). New York: ACM.

Saltman, K. (2020). Artificial intelligence and the technological turn of public education privatization. *London Review of Education, 18*(2), 196–208.

Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement*, 143–152.

Shaffer, D. W. (2006a). Epistemic frames for epistemic games. *Computers in Education, 46*(3), 223–234.

Shaffer, D. W. (2006b). *How computer games help children learn*. New York, NY: Palgrave.

Shaffer, D. W., Collier, W., & Ruis, A. R. (2016). A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *Journal of Learning Analytics, 3*(3), 9–45.

Shaffer, D. W., & Kaput, J. J. (1998). Mathematics and virtual culture: An evolutionary perspective on technology and mathematics education. *Educational Studies in Mathematics, 37*, 97–119.

Shin, D., Shim, Y., Yu, H., Lee, S., Kim, B., & Choi, Y. (2021). Saint+: Integrating temporal features for ednet correctness prediction. In *Proceedings of the 11th international learning analytics and knowledge conference* (pp. 490–496).

Shnayder, V., & Parkes, D. C. (2016). Practical peer prediction for peer assessment. In *Proceedings of the fourth AAAI conference on human computation and crowdsourcing* (pp. 199–208).

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer games and instruction, 55*(2), 503–524.

Shute, V. J., & Rahimi, S. (2021). Stealth assessment of creativity in a physics video game. *Computers in Human Behavior, 116*, Article 106647.

Shute, V., Rahimi, S., Smith, G., Ke, F., Almond, R., Dai, C. P., … Sun, C. (2021). Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity and learning supports in educational games. *Journal of Computer Assisted Learning, 37*(1), 127–141.

Shute, V., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. Cambridge, MA: MIT Press.

Sorrel, M. A., Barrada, J. R., de la Torre, J., & Abad, F. J. (2020). Adapting cognitive diagnosis computerized adaptive testing item selection rules to traditional item response theory. *PLoS One, 15*(1), Article e0227196.

Sullivan, S. A., Warner-Hillard, C., Eagan, B., Thompson, R., Ruis, A. R., Haines, K., et al. (2018). Using epistemic network analysis to identify targets for educational interventions in trauma team communication. *Surgery, 163*(4), 938–943.

Suto, I., Nádas, R., & Bell, J. (2011). Who should mark what? A study of factors affecting marking accuracy in a biology examination. *Research Papers in Education, 26*(1), 21–51.

Swauger, S. (2020). Our bodies encoded: Algorithmic test proctoring in higher education. In J. Stommel, C. Friend, & S. Morris (Eds.), *Critical digital pedagogy*. Press Books.

Taras, M. (2008). Assessment for learning. *Journal of Further and Higher Education, 32*(4), 389–397.

Thompson, N. A. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment, Research and Evaluation, 12*(1), 1.

Topping, K. J. (2018). *Using peer assessment to inspire reflection and learning*. Routledge.

Toton, S. L., & Maynes, D. D. (2019). Detecting examinees with pre-knowledge inExperimental data using conditional scaling of response times. *Frontiers in Education, 4*, 49.

Van Der Graaf, J., Lim, L., Fan, Y., Kilgour, J., Moore, J., Bannert, M., … Molenaar, I. (2021). April). Do instrumentation tools capture self-regulated learning?. In *LAK21: 11th international learning analytics and knowledge conference* (pp. 438–448).

Verschoor, A., Berger, S., Moser, U., & Kleintjes, F. (2019). On-the-Fly calibration in computerized adaptive testing. In *Theoretical and practical advances in computer-based educational measurement* (pp. 307–323). Cham: Springer.

Vygotsky, L. S., & Cole, M. (1978). *Mind in society: Development of higher psychological processes*. Harvard University Press.

Wang, W., An, B., & Jiang, Y. (2018). Optimal spot-checking for improving evaluation accuracy of peer grading systems. In *Proceedings of the 32nd AAAI conference on artificial intelligence* (pp. 833–840).

Whitehill, J., Aguerrebere, C., & Hylak, B. (2019). Do learners know what's good for them? Crowdsourcing subjective ratings of oers to predict learning gains. In *Proceedings of the 12th international conference on educational data mining* (pp. 462–467). IEDMS.

Wiliam, D. (2011). What is assessment for learning? *Studies In Educational Evaluation, 37*(1), 3–14.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. New York: Taylor & Francis Group.

Wilson, J., Ahrendt, C., Fudge, E. A., Raiche, A., Beard, G., & MacArthur, C. (2021). Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation. *Computers & Education, 168*, Article 104208.

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education, 100*, 94–109.

Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research, 58*(1), 87–125.

Wilson, M., & Scalise, K. (2012). Assessment of learning in digital networks. In P. Griffin, & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 37–56). Dordrecht: Springer.

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education, 30*(4), 343–354.

Wright, J. R., Thornton, C., & Leyton-Brown, K. (2015). Mechanical TA: Partially automated high-stakes peer grading. In *Proceedings of the 46thACM technical symposium on computer science education* (pp. 96–101).

Yannakoudakis, H., & Briscoe, T. (2012). Modeling coherence in ESOL learner texts. In *Proceedings of the seventh workshop on building educational applications using NLP* (pp. 33–43).

Zheng, Y., Li, G., Li, Y., Shan, C., & Cheng, R. (2017). Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment, 10*(5), 541–552.

Zhou, M., & Winne, P. H. (2012). Modeling academic achievement by self-reported versus traced goal orientation. *Learning and Instruction, 22*(6), 413–419.